The Thirty-Fourth AAAI Conference on Artiﬁcial Intelligence (AAAI-20)

**Differentially Private and Fair Classiﬁcation via Calibrated Functional Mechanism**

# Jiahao Ding,1 Xinyue Zhang,1 Xiaohuan Li,2 Junyi Wang,2 Rong Yu,3 Miao Pan1

1University of Houston

2Guilin University of Electronic Technology 3Guangdong University of Technology

{jding7, xzhang67, mpan2}@uh.edu, {lxhguet, wangjy}@guet.edu.cn, [yurong@gdut.edu.cn](mailto:yurong@gdut.edu.cn)

### Abstract

Machine learning is increasingly becoming a powerful tool to make decisions in a wide variety of applications, such as medical diagnosis and autonomous driving. Privacy concerns related to the training data and unfair behaviors of some deci- sions with regard to certain attributes (e.g., sex, race) are be- coming more critical. Thus, constructing a fair machine learn- ing model while simultaneously providing privacy protection becomes a challenging problem. In this paper, we focus on the design of classiﬁcation model with fairness and differen- tial privacy guarantees by jointly combining functional mech- anism and decision boundary fairness. In order to enforce *g*- differential privacy and fairness, we leverage the functional mechanism to add different amounts of Laplace noise re- garding different attributes to the polynomial coefﬁcients of the objective function in consideration of fairness constraint. We further propose an utility-enhancement scheme, called re-

laxed functional mechanism by adding Gaussian noise in- stead of Laplace noise, hence achieving (*g, δ*)-differential privacy. Based on the relaxed functional mechanism, we can design (*g, δ*)-differentially private and fair classiﬁcation model. Moreover, our theoretical analysis and empirical re- sults demonstrate that our two approaches achieve both fair- ness and differential privacy while preserving good utility and outperform the state-of-the-art algorithms.

# Introduction

In this big data era, machine learning has been becoming a powerful technique for automated and data-driven decision making processes in various domains, such as spam ﬁltering, credit ratings, housing allocation, and so on. However, as the success of machine learning mainly rely on a vast amount of individual data (e.g., ﬁnancial transactions, tax payments), there are growing concerns about the potential for privacy leakage and unfairness in training and deploying machine learning algorithms (Fredrikson, Jha, and Ristenpart 2015; Datta, Tschantz, and Datta 2015). Thus, the problem of fair- ness and privacy in machine learning has attracted consider- able attention.

Fairness-aware learning has received growing attentions in the machine learning ﬁeld due to the social inequities and

Copyright © 2020, Association for the Advancement of Artiﬁcial Intelligence (www.aaai.org). All rights reserved.

unfair behaviors observed in classiﬁcation models. For ex- ample, a classiﬁcation model of automated job hiring system is more likely to hire candidates from certain racial or gen- der groups (Giang 2018; Wachter-Boettcher 2018). Hence, substantial effort has centered on developing algorithmic methods for designing fair classiﬁcation models and bal- ancing the trade-off between accuracy and fairness, mainly including two groups: pre/post-processing methods (Dwork et al. 2012; Feldman et al. 2015; Hardt et al. 2016) and in- processing methods (Kamishima, Akaho, and Sakuma 2011; Zafar et al. 2017b). Pre/post-processing methods achieve fairness by directly changing values of the sensitive at- tributes or class labels in the training data. As pointed out in (Zafar et al. 2017b), pre/post-processing methods treat the learning algorithm as a black box, which can result in unpre- dictable loss of the classiﬁcation utility. Thus, in-processing methods, which introduce fairness constraints or regulariza- tion terms to the objective function to remove the discrimi- natory effect of classiﬁers, have been shown a great success.

At the same time, differential privacy (Dwork and Roth 2014) has emerged as the de facto standard for measuring the privacy leakage associated with algorithms on sensitive databases, which has recently received considerable atten- tions by large-scale corporations such as Google (Erlings- son, Pihur, and Korolova 2014) and Microsoft (Ding, Kulka- rni, and Yekhanin 2017), etc. Generally speaking, differ- ential privacy ensures that there is no statistical difference to the output of a randomized algorithm whether a single individual opts in to, or out of its input. A large class of mechanisms has been proposed to ensure differential pri- vacy. For instance, the Laplace mechanism is employed by introducing random noise drawn from the Laplace distribu- tion to the output of queries such that the adversary will not be able to conﬁrm a single individual is in the input with high conﬁdence (Dwork et al. 2006b). To design private machine learning models, more complicated perturbation mechanisms have been proposed like objective perturbation (Chaudhuri, Monteleoni, and Sarwate 2011) and functional mechanism (Zhang et al. 2012), which inject random noise into the objective function rather than model parameters.

Thus, in this paper, we mainly focus on achieving classi- ﬁcation models that simultaneously provide differential pri-

vacy and fairness. As pointed out in recent study (Xu, Yuan, and Wu 2019), achieving both requirements efﬁciently is quite challenging, due to the different aims of differential privacy and fairness. Differential privacy in a classiﬁcation model focuses on the individual level, i.e., differential pri- vacy guarantees that the model output is independent of whether any individual record presents or absents in the dataset, while fairness in a classiﬁcation model focuses on the group level, i.e., fairness guarantees that the model pre- dictions of the protected group (such as female group) are same to those of the unprotected group (such as male group). Lots of researches have emerged in achieving both privacy protection and fairness. Speciﬁcally, in (Dwork et al. 2012), Dwork et al. gave a new deﬁnition of fairness that is an extended deﬁnition of differential privacy. In (Hajian et al. 2015), Hajian et al. imposed fairness and *k*-anonymity via a pattern sanitization method. Moreover, Ekstrand et al. in (Ekstrand, Joshaghani, and Mehrpouyan 2018) put forward a set of questions about whether fairness are compatible with privacy. However, only Xu et al. in (Xu, Yuan, and Wu 2019) studied how to meet the requirements of both differential privacy and fairness in classiﬁcation models by combining

functional mechanism and decision boundary fairness to- gether. Therefore, how to simultaneously meet the require-

proposing the relaxed functional mechanism based on Ex- tended Gaussian mechanism, and leverage it to introduce Gaussian noise with different scales to perturb objective function.

* Using real-world datasets, we show that the performance of PDFC and ADFC signiﬁcantly outperforms the base- line algorithms while jointly providing differential pri- vacy and fairness.

The rest of paper is organized as follows. We ﬁrst give the problem statement and background in differential privacy and fairness. Next, we present our two approaches PDFC and ADFC to achieve DP and fair classiﬁcation. Finally, we give the numerical experiments based on real-world datasets and draw conclusion remarks. Due to the space limit, we leave all the proofs in the supplemental materials.

# Problem Statement

This paper considers a training dataset *D* that includes *n* tu- ples *t*1*, t*2*, , tn*. We also denote each tuple *ti* = (***x****i, yi*) where the feature vector ***x****i* contains *d* attributes, i.e., ***x****i* = (*xi*1*, xi*2*,* ··· *, xid*), and *yi* is the co.rresponding label. With-

···

Σ

out loss of generality, we assume

*d*

*j*=1

*x*

2

*ij*

≤ 1 where

ments of differential privacy and fairness in machine learn-

ing algorithms is under exploited.

In this paper, we propose **P**urely and **A**pproximately **D**ifferential private and **F**air **C**lassiﬁcation algorithms, called PDFC and ADFC, respectively, by incorporating functional mechanism and decision boundary covariance, a novel measure of decision boundary fairness. As shown in (Kamiran and Calders 2012), due to the correlation between input features (attributes), the discrimination of classiﬁca-

tion still exists even if removing the protected attribute from

*xij* 0, and *yi* 0*,* 1 for binary classiﬁcation tasks.

The objective is to construct a binary classiﬁcation model *ρ*(***x****, w*) with model parameters *w* = (*w*1*, w*2*, , wd*) that taken ***x*** as input, can output the prediction *y*ˆ, by minimizing the empirical loss on the training dataset *D* over the param- eter space *w* of *ρ*.

···

≥ ∈ { }

In general, we have the following optimization problem.

*w*∗ = arg min *f* (*D, w*) = arg min Σ *f* (*ti, w*) (1)

*n*

*w*

*w*

*i*=1

the dataset before training. Hence, different from (Xu, Yuan,

and Wu 2019), which adds same scale of noise in each at- tribute, in PDFC, we consider a calibrated functional mech-

anism, i.e., injecting different amounts of Laplace noise re-

where *f* is the loss function. In this paper, we consider logistic regression as the loss function, i.e., *f* (*D, w*) =

Σ*n* [log(1+*exp*(***x****T w*))−*yi****x****T w*]. Thus, the classiﬁcation

*i*=1

*i*

*i*

garding different attributes to the polynomial coefﬁcients of

model has the form *ρ*(***x****, w*∗) = *exp*(***x****T w*∗)

.

the constrained objective function to ensure *s*-differential

1+*exp*(***x****T w*∗)

Although there is no need to share the dataset during the

privacy and reduce effects of discrimination. To further im- prove the model accuracy, in ADFC, we propose a relaxed functional mechanism by inserting Gaussian noise instead of Laplace noise and leverage it to perturb coefﬁcients of the polynomial representation of the constrained objective function to enforce (*s, δ*)-differential privacy and fairness. Our salient contributions are listed as follows.

* We propose two approaches PDFC and ADFC to learn a logistic regression model with differential privacy and fairness guarantees by applying functional mechanism to a constrained objective function of logistic regression that decision boundary fairness constraint is treated as a penalty term and added to the original objective function.
* For PDFC, different magnitudes of Laplace noise regard- ing different attributes are added to the polynomial coef- ﬁcients of the constrained objective function to enforce *s*-differential privacy and fairness.
* For ADFC, we further improve the model accuracy by

training procedure, the risk of information leakage still ex- ists when we release the classiﬁcation model parameter *w*∗. For example, the adversary may perform model inversion attack (Fredrikson, Jha, and Ristenpart 2015) over the re-

lease model *w*∗ together with some background knowledge about the training dataset to infer sensitive information in the dataset.

Furthermore, if labels in the training dataset are associ- ated with a protected attribute *zi* (note that we denote ***x****i* as unprotected attributes), like gender, the classiﬁer may be bi- ased, i.e., *P* (*y*ˆ*i* = 1 *zi* = 0) = *P* (*y*ˆ*i* = 1 *zi* = 1), where we assume the protected attribute *zi* 0*,* 1 . According to (Pedreshi, Ruggieri, and Turini 2008), even if the protected attribute is not used to build the classiﬁcation model, this unfair behavior may happen when the protected attribute is correlated with other unprotected attributes.

∈ { }

| ƒ |

Therefore, in this paper, our objective is to learn a binary classiﬁcation model, which is able to guarantee differential privacy and fairness while preserving good model utility.

# Background

In this section, we ﬁrst introduce some background knowl- edge of differential privacy, which helps us to build private classiﬁcation models. Then we present fairness deﬁnition, which helps us to enforce classiﬁcation fairness.

## Differential Privacy

Differential privacy is introduced to guarantee that the abil- ity of an adversary to obtain additional information about any individual is independent of whether any individual

mechanism is designed for regression analysis. To preserve *s*-differential privacy, functional mechanism injects differ- entially private noise into the objective function *f* (*D, w*) and then publishs a noisy model parameter *w*ˆ derived from minimizing the perturbed objective function *f*ˆ(*D , w*) rather than the original one. As a result of the objective function being a complex function of *w*, in functional mechanism, *f* (*D, w*) is represented in polynomial forms trough Taylor Expansion. The model parameter *w* is a vector consisting of several values *w*1*, w*2*,* ··· *, wd*. We denote *φ*(*w*) as a prod-

uct of *w*1*, w*2*,* ··· *, wd*, namely, *φ*(*w*) = *wc*1 *wc*2 ··· *wcd* for

record presents or absents in the dataset.

**Deﬁnition 1** (***s*-Differential Privacy).** *A randomized Mech-*

some *c*1*, c*2*,* ··· *, cd* ∈ N

1 2 *d*

. We also denote Φ*j*(*j* ∈ N) as

the set of all products of *w*1*, w*2*,* ··· *, wd* with degree *j*, i.e.,

1

2

*d*

*l*=1

*anism* A *is enforced by s-differential privacy, if for any two*

*neighboring datasets D, D* ∗ ∈ D*, i.e., differing at most*

Φ*j* = {*wc*1 *wc*2 ··· *wcd* | Σ*d cl* = *j*}.

*one single data sample, and for any possible output s in the output space of* A*, it holds that* Pr(A(*D* ) = *s*) ≤ *es* Pr(A(*D* ∗) = *s*)*.*

The privacy parameter *s* controls the strength of the pri-

According to the Stone-Weierstrass Theorem (Rudin and others 1964), any continuous and differentiable function can always be expressed as a polynomial form. Therefore, the objective function *f* (*D, w*) can be written as follows

*n J*

Σ Σ Σ

vacy guarantee. A smaller value indicates a stronger privacy protection. Though differential privacy provides very strong guarantee, in some cases it may be too strong to have a good data utility. We then introduce a relaxation, (*s*,*δ*)-differential privacy, that has been proposed in (Dwork et al. 2006a).

**Deﬁnition 2** ((***s***,***δ***)**-Differential Privacy).** *A randomized*

*f* (*D, w*) = *λφti φ*(*w*)*,* (2)

*i*=1 *j*=0 *φ*∈Φ*j*

where *λφti* represents the coefﬁcient of *φ*(*w*) in polynomial. To preserve *s*-differential privacy, the objective function  *f* (*D, w*) is perturbed by adding Laplace noise into the poly-

nomial coefﬁcients, i.e., *λφ* = Σ*n*

where Δ1 = 2 max*t*

*j*=1

*λφt*

+ *Lap*(Δ1*/s*),

*Mechanism* A *is enforced by (s,δ)-differential privacy, if*

*for any two neighboring datasets D, D* ∗ ∈ D *differing at*

*φ*∈Φ*j* ǁ*λφt*ǁ1. And then the

Σ*J* Σ

*i*=1 *i*

*most one single data item, and for any possible output s*

*in the output space of* A*, it holds that Pr*(A(*D* ) = *s*) ≤

*esPr*(A(*D* ∗) = *s*)+ *δ.*

Laplace mechanism (Dwork and Roth 2014) and Ex- tended Gaussian mechanism (Phan et al. 2019) are common techniques for achieving differential privacy, both of which add random noise calibrated to the sensitivity of the query

model parameter *w*ˆ is obtained by minimizing the noisy ob-

jective function *f*ˆ(*D , w*). The sensitivity of logistic regres-

sion is given in the following lemma

**Lemma 1** (*l*1**-Sensitivity of Logistic Regression).** *Let f* (*D, w*) *and f* (*D* ∗*, w*) *be the logistic regres- sion on two neighboring datasets D and D* ∗*, respec-*

*tively, and denote their polynomial representations*

*i*=1

Σ

Σ

function *q*.

*as f* (*D, w*) = Σ*n*

*i*=1

*J*

*j*=1

Σ*φ*∈Φ*j*

*i*

*λφti φ*(*w*) *and*

**Theorem 1** (**Laplace Mechanism).** *Given any function q* :

X *n* → R*d, the Laplace mechanism deﬁned by*

2

*f* (*D* ∗*, w*) = Σ*n*

*J*

*j*=1

*have the following inequality*

Σ*φ*∈Φ*j*

*λφt*∗ *φ*(*w*)*. Then, we*

M*L*(*D, q, s*) = *q*(*D* )+ (*Y*1*, Y*2*,* ··· *, Yd*)

*preserves s-differential privacy, where Yi are i.i.d. random*

Δ1 = Σ

Σ ǁ Σ

*λφti* −

Σ *λφt*∗*i* ǁ1

*variables drawn from Lap*(Δ1*q/s*) *and l*1*-sensitivity of the query q is* Δ1*q* = sup*D,D*∗ ǁ*q*(*D* ) − *q*(*D* ∗)ǁ1 *taken over all neighboring datasets D and D* ∗*.*

*j*=1 *φ*∈Φ*j*

2

Σ≤

2 max

*t*

*ti*∈*D*

Σ

*t*∗*i*∈*D*∗

*d*2

ǁ*λφt*ǁ1 ≤ 4

+ *d,*

**Theorem 2** (**Extended Gaussian Mechanism).** *Given any function q* : *n* R*d and for any s >* 0*, δ* (0*,* 1)*, the Extended Gaussian mechanism deﬁned by*

X → ∈

M*G*(*D, q, s*) = *q*(*D* )+ (*Y*1*, Y*2*,* ··· *, Yd*)

*preserves* (*s, δ*)*-differential privacy, where Yi are i.i.d*

*drawn f*.*rom a Gaussian*.*distribution* N (0*, σ*2*Id*) *with σ* ≥

*j*=1 *φ*∈Φ*j*

*where ti, t*∗*i or t is an arbitrary tuple.*

## Classiﬁcation Fairness

The goal of classiﬁcation fairness is to ﬁnd a classiﬁer that minimizes the empirical loss while guaranteeing certain fair- ness requirements. Many fairness deﬁnitions have been pro- posed for in the literature including mistreatment parity (Za-

√2Δ2*q* (

2*s*

*π δ*

*π δ*

2

log(. 2 1 )+

log(. 2 1 )+ *s*) *and l -sensitivity*

far et al. 2017a), demographic parity (Pedreshi, Ruggieri,

*of the query q is* Δ2*q* = sup*D,D*

*over all neighboring datasets*

∗ ǁ*q*(*D* ) − *q*(*D* ∗)ǁ2 *taken*

∗*.*

and Turini 2008), etc.

Demographic parity, the most widely-used fairness deﬁni-

*D and D*

**Functional Mechanism.** Functional mechanism, intro- duced by (Zhang et al. 2012), as an extension of the Laplace

tion in the classiﬁcation fairness domain, requires the deci- sion made by the classiﬁer is not dependent on the protected attribute *z*, for instance, sex or race.

**Deﬁnition 3.** *(***Demographic Parity in a Classiﬁer)** *Given a classiﬁcation model y*ˆ = *ρ*(***x****, w*) *and a labeled dataset D, the property of demographic parity in a classiﬁer is deﬁned by* Pr(*y*ˆ = 1 *z* = 1) = Pr(*y*ˆ = 1 *z* = 0) *where z* 0*,* 1 *is the protected attribute.*

| | ∈ { }

Moreover, demographic parity is quantiﬁed in terms of the risk difference (RD) (Pedreschi, Ruggieri, and Turini

suitable values to make *α*1 = 1. Note that our theoretical results still hold if we choose other values of *α*1 and *τ* . By equation (4), we have

*n*

*f*˜(*D , w*) = [log(1 + *exp*(***x****T w*)) − *yi****x****T w*]

*i i*

Σ

*i*=1

.Σ .

*n*

.

+

*i*

2012), i.e., the difference of the positive decision made in

between the protected group and unprotected group. Thus,

.*i*=1

(*zi* − *z*¯)***x****T w*. *.* (5)

the risk difference produced by a classiﬁer is deﬁned as

*RD* = Pr(*y*ˆ = 1 *z* = 1) Pr(*y*ˆ = 1 *z* = 0) *.*

| | − | |

One of the in-processing methods, called decision bound-

To apply functional mechanism, we ﬁrst write the approxi- mate objective function *f*¯(*D , w*) based on (2) as follows.

ary fairness (Zafar et al. 2017b), to ensure classiﬁcation fair- ness is to ﬁnd a model parameter *w* that minimizes the loss

1

*f*¯(*D , w*) = Σ Σ

*f* (*j*)(0)

(***x****T w*)*j* −

.Σ*n*

*yi****x****T* Σ *w*

function *f* (*D*

*, w*) under a fairness constraint. Thus, the fair

*j*! *i*

*i*=1 *j*=0

*n*

2

...Σ

*i*

*i*=1

*i*

classiﬁcation problem is formulated as follows,

*minimize f* (*D, w*)

*subject to g*(*D, w*) ≤ *τ, g*(*D, w*) ≥ −*τ,* (3)

*n*

+

*i*=1

Σ Σ Σ

2

*n*

(*zi* − *z*¯)***x****T w*.

where *g*(*D, w*) is a constraint term, and *τ* is the threshold. For instance, Zafar et al. (Zafar et al. 2017b) have proposed to adopt the decision boundary covariance to deﬁne the fair-

where *λ*¯*φt*

.

= *λ*¯*φti φ*(*w*)*,* (6)

*i*=1 *j*=0 *φ*∈Φ*j*

denotes the coefﬁcient of *φ*(*w*) in the polynomial

ness constraint, i.e.,

*g*(*D, w*) = E[(*z* − *z*¯)*d*(***x****, w*)] − E[*z* − *z*¯]*d*(***x****, w*)

Σ

*n*

∝ (*zi* − *z*¯)*d*(***x****i, w*)*,* (4)

*i*=1

*i*

of *f*¯(*ti, w*) and *f*1( ) = log(1 + exp ( )).

· ·

The attributes involving in the dataset may not be inde- pendent from each other, which means some unprotected attributes in ***x*** are quite correlated with the protected at- tribute *z*. For instance, the protected attribute, like gender, may be correlated with the attribute, marital status. Thus, to

where {*d*(***x****i, w*)}*n* is decision boundary, *z*¯ is the average

*i*=1

reduce the discrimination between the protected attribute *z*

of the protected attribute and E[*z z*¯] = 0. For logistic re- gression classiﬁcation models, the decision boundary is de- ﬁned by ***x****T w*. The decision boundary covariance (4) then

−

and the labels *y*, it is important to weaken the correlation between these most correlated attributes and protected at- tribute *z*. However, it is often impossible to determine the

reduces to *g*(*D, w*) = Σ*n* (*zi* − *z*¯)***x****T w*.

*i*=1

*i*

degree of relation between an unprotected attribute and the

protected attribute. Therefore, we randomly select an unpro-

# Differentially Private and Fair Classiﬁcation

In this section, we ﬁrst present our approach PDFC to achieve fair logistic regression with *s*-differentially private guarantee. Then we propose a relaxed functional mecha- nism by injecting Gaussian noise instead of Laplace noise to provide (*s, δ*)-differential privacy. By leveraging the re- laxed functional mechanism, we will show that our second approach ADFC can jointly provide (*s, δ*)-differential pri- vacy and fairness.

## Purely DP and Fair Classiﬁcation

In order to meet the requirements of *s*-differential privacy and fairness, motivated by (Xu, Yuan, and Wu 2019), we consider to combine the functional mechanism and decision boundary fairness. We ﬁrst consider to transform the con- strained optimization problem (3) into unconstrained prob- lem by treating the fairness constraint as a penalty term, where the fairness constraints are shifted to the original objective function *f* (*D, w*). Then, we have the new ob-

jective function *f*˜ (*w*) deﬁned as *f*˜(*D , w*) = *f* (*D, w*) +

tected attribute *xs* and leverage functional mechanism to add noise with large scale to the corresponding polynomial co- efﬁcients of the monomials involving *ws*. Interestingly, this approach not only helps to reduce the correlation between attributes *xs* and *z*, but also improve the privacy on attribute *xs* to prevent model inversion attacks, as shown in (Wang, Si, and Wu 2015).

The key steps of PDFC are outlined in Algorithm 1. We ﬁrst set two different privacy budgets, *ss* and *sn*, for attribute *xs* and the rest of attributes ***x*** *xs* . Before injecting noise to the coefﬁcients, all coefﬁcients *φ* should be separated into two groups Φ*s* and Φ*n* by considering whether *ws* in- volves in the corresponding monomials (i.e., whether their the coefﬁcients contain attribute *xs*). We then add Laplace noises drawn from *Lap*(Δ1*/ss*) and *Lap*(Δ1*/sn*) to the co- efﬁcients of *φ* Φ*s* and *φ* Φ*n* respectively to reconstruct the differentially private objective function *f*ˆ(*D , w*), where Δ1 can be found in Lemma 2. Finally, the differentially pri- vate model parameter *w*ˆ is obtained by minimizing *f*ˆ(*D , w*). Note that *w*ˆ also ensures classiﬁcation fairness due to the ob-

∈ ∈

{ \ }

*α*1|*g*(*D*

*, w*) − *τ* |

*D*

, where we consider *α*1

as a hyperparameter

jective function involving fairness constraint.

to optimize the trade-off between model utility and fairness. For convenience of discussion, we set *τ* = 0 and choose

**Lemma 2.** *Let D and D* ∗ *be any two neighboring datasets differing in at most one tuple. Let f*¯(*D , w*) *and f*¯(*D* ∗*, w*) *be*

**Algorithm 1** Purely DP and Fair Classiﬁcation (PDFC)

1: **Input:** Dataset *D* ; The objective function *f* (*D, w*); The

**Algorithm 2** Relaxed Functional Mechanism

1: **Input:** Dataset *D* ; The objective function *f* (*D, w*) =

*φ*∈Φ*j*

*i*

fairness constraint *g*(*D, w*); The privacy budget *ss* for

unprotected attribute *x* ; The privacy budget *s*

for other

Σ*n* Σ*J* Σ

*i*=1

*j*=1

*λφt φ*(*w*); The privacy parameters

*s*

{

unprotected attributes ***x***

2: **Output:** *w*ˆ, *s*.

\ *xs*};

¯

*n*

*l*1-sensitivity Δ1.

*s, δ*.

2: **Output:** *w*ˆ

3: Set Δ2 according Lemma 3.

3: Set the approximate function *f* (*D, w*) by equation (6).

4: Set two sets Φ*s* = {}, Φ*n* = {}.

5: **for** 1 ≤ *j* ≤ 2 **do**

4: **for** 1 ≤ *j* ≤ *J* **do**

5: **for** each *φ* ∈ ΦΣ*j* **do**

.

6: Set *λφ* =

*n*

*λφt*

+ N (0*, σ*2), where *σ* =

.

.

.

(

)+

6: **for** each *φ* ∈ Φ*j* **do**

*i*=1 *i*

7: **if** *φ* includes *ws* for a particular attribute *xs*

### then

8: Put *φ* into Φ*s*.

9: **else**

√2Δ2

2*s*

*π δ*

*π δ*

7: **end for**

8: **end for**

log( 2 1

log(

2 1 )+ *s*).

10: Put *φ* into Φ*n*.

9: Let *f*ˆ(*D , w*) = Σ*J*

Σ*φ*∈Φ*j*

*λφφ*(*w*).

11: **end if**

12: **end for**

13: **end for**

14: **for** 1 ≤ *j* ≤ 2 **do**

15: **for** each *φ* ∈ Φ*j* **do**

16: **if** *φ* ∈ Φ*s* **then**

Σ*n*ˆ17: Set *λ* =*φ*

*i*=1

18: **else**

Σ

19: Set *λ*ˆ*φ* = *n*

20: **end if**

21: **end for**

*i*=1

*λ*¯*φti λ*¯*φt*

+ *Lap*(Δ1*/*(*ss*)).

+ *Lap*(Δ1*/*(*sn*)).

10: Compute *w*ˆ = arg min*w f*ˆ(*D , w*).

11: **return:** *w*ˆ.

*j*=1

**Lemma 3** (*l*2**-Sensitivity of Logistic Regression).** *For polynomial representations of logistic regression, two f* (*D, w*) *and f* (*D* ∗*, w*) *given in Lemma 1, we have the fol- lowing inequality*

.

*i*

*i*=1

*φ*∈∪*j*=1Φ*j*

Δ2 = ǁ*A*1 − *A*2ǁ2 ≤

*d*2

+ *d,*

16

22: **end for**

Let ˆ

2

Σ Σ

.

24: Compute *w*ˆ = arg min

*f*ˆ(*D , w*).

*where we denote A*1 = {Σ*n*

ˆ

*i*

*i*=1

*φti*

*φ*∈∪*J*

Φ*j*

*λφt* } *J*

*and A*2 =

23:

*f* (*D, w*) =

*j*=1

*φ*∈Φ*j λφφ*(*w*)

.Σ*n*

*λ* ∗ Σ

*as the set of polynomial coefﬁcients*

*w*

25: Compute *s* = *ss/d* + *sn*(*d* − 1)*/d.*

*of f* (*D, w*)

*and*

*j*=1

*f* (*D* ∗*, w*)

*. And we denote ti*

*or t*∗*i as an ar-*

26: **return:** *w*ˆ, *s*.

*bitrary tuple.*

We then perturb *f* (*D, w*) by injecting Gaussian noise

.

2*s*

*π δ*

drawn from N (0*, σ*2) with *σ* = √2Δ2 ( log(. 2 1 ) +

*have the following inequality,*

*π δ*

*the approximate objective function on D and D* ∗*, then we*

.log(. 2 1 )+ *s*) into its polynomial coefﬁcients, and ob-

2 *n n* 2

Δ = Σ Σ ǁ Σ *λ*¯ − Σ *λ*¯ ∗ ǁ ≤ *d*

+ 3*d.*

tain the differentially private model parameter *w*ˆ by mini-

1

*d*

*d*

*j*=1 *φ*∈Φ*j*

*i*=1

*φti*

*i*=1

*φti* 1 4

mizing the noisy function *f*ˆ(*D , w*), as shown in Algorithm

2. Finally, we provide a privacy guarantee of proposed re-

The following theorem shows the privacy guarantee of

Σ

function of logistic regression *f* (*D, w*) =

*n*

*i*=1

PDFC.

**Theorem 3.** *The output model parameter w*ˆ *in PDFC (Al-*

*gorithm 1) preserves s-differential privacy, where s* = 1 *ss* +

*d*−1 *sn.*

## Approximately DP and Fair Classiﬁcation

We now focus on using the relaxed version of *s*-differential privacy, i.e., (*s, δ*)-differential privacy to further improve the utility of differentially private and fair logistic regression. Hence, in order to satisfy (*s, δ*)-differential privacy, we pro- pose the relaxed functional mechanism by making use of Extended Gaussian mechanism. As shown in Theorem 2, before applying Extended Gaussian mechanism, we ﬁrst cal- culate the sensitivity of a query function, i.e., the objective

[log(1 +

laxed functional mechanism by the following theorem.

**Theorem 4.** *The relaxed functional mechanism in Algo-*

{ \ }

to polynomial coefﬁcients of *φ* ∈ Φ

*s s*. For polynomial coef-

*rithm 2 guarantees* (*s, δ*)*-differential privacy.*

Our second approach called, ADFC, applies the re-

laxed functional mechanism into the objective function

with decision boundary fairness constraint to enforce (*s, δ*)-

differential privacy and fairness. As shown in Algorithm 3,

we ﬁrst derive the polynomial representation *f*¯(*D , w*) ac-

cording to (6), and employ random Gaussian noise to per- turb the objective function *f*¯(*D , w*), i.e., injecting Gaussian noise into its polynomial coefﬁcients. Furthermore, we also

allocate differential privacy parameters, (*ss, δs*) and (*sn, δn*) for a particular unprotected attribute *xs* and the rest of un- protected attributes ***x*** *xs* to improve the privacy on at- tribute *xs* and reduce the correlation between attributes *xs* and *z*. Hence, we add random noise drawn from N (0*, σ*2)

*exp*(***x****T w*)) − *yi****x****T w*], given in the following lemma.

ﬁcients in Φ*n*, we inject noise drawn from N (0*, σ*2 ).

*i*

*i*

*n*

**Algorithm 3** Approximately DP and Fair Classiﬁcation (ADFC)

1: **Input:** Dataset *D* ; The objective function *f* (*D, w*); The fairness constraint *g*(*D, w*); The privacy parame- ters *ss, δs* for unprotected attribute *xs*; The privacy pa- rameters *sn, δn* for other unprotected attributes ***x*** *xs* .

{ \ }

2: **Output:** *w*ˆ, *s* and *δ*.

3: Set the approximate function *f*¯(*D , w*) by equation (6).

**0.74**

**0.72**



**)XQFWLRQDO 0HFKDQLVP 5HOD[HG**

**)XQFWLRQDO 0HFKDQLVP 1R-3ULYDF\**

**0.7**

**0.68**

**$FFXUDF\**

**0.66**

**0.64**

**0.62**

**0.6**

4: Set two sets Φ*s* = {}, Φ*n* = {}.

**10-2 10-1 100 101**

**3ULYDF\ %XGJHW**

5: **for** 1 *j* 2 **do**

≤ ≤

6: **for** each *φ* Φ*j* **do**

∈

7: **if** *φ* includes *ws* for a particular attribute *xs*

### then

8: Put *φ* into Φ*s*.

9: **else**

10: Put *φ* into Φ*n*.

11: **end if**

12: **end for**

13: **end for**

14: Set *l*2-sensitivity Δ∗2 by Lemma 4.

15: **for** 1 ≤ *j* ≤ 2 **do**

16: **for** each *φ* ∈ Φ*j* **do**

17: **if** *φ* ∈ Φ*s* **then**

Figure 1: Compare accuracy under different privacy budgets on *US.* (*δ* = 10−3)

**0.74**

**)XQFWLRQDO 0HFKDQLVP 5HOD[HG**

**)XQFWLRQDO 0HFKDQLVP**

**1R-3ULYDF\**

**0.72**

**0.7**

**$FFXUDF\**

**0.68**

**0.66**

**0.64**

**0.62**

**10-3 10-4 10-5 10-6 10-7**

18: Set *λ*ˆ*φ* = Σ*n λ*¯*φt*  + N (0*, σ*2),

√2Δ∗

2*ss*

where *σ*

= (

log(

) +

Figure 2: Compare accuracy under different values of *δ* on

19: **else**

*i*=1 .*i* . *s*

.log(. 2 1 )+ *ss*).

2

*π δs*

*s*

2 1

*π δs*

*US*.

function *f*ˆ(*D , w*), we derive the model parameter *w*ˆ, which achieves differential privacy and fairness at the same time.

20: Set

*λ*ˆ*φ* = Σ*n λ*¯*φt*  + N (0*, σ*2 ),

√2Δ∗

= (

We now show that ADFC satisﬁes (*s, δ*)-differential privacy

*i*=1 .*i* . *n*

in the following theroem.

2

where *σ*

log(

2

1 ) +

log(

2 1 )+ *sn*).

*n* 2*sn*

. .

*π δn*

**Theorem 5.** *The output model parameter w*ˆ *in ADFC (Al-*

*gorithm 3) guarantees* (*s, δ*)*-differential privacy, where s* =

21: **end if**

22: **end for**

23: **end for**

Σ24: Let *f* (*D, w*) =

ˆ 2

*j*=1

*π δn*

Σ*φ*∈Φ*j*

*λ*ˆ*φφ*(*w*).

*d d*

# Performance Evaluation

1 *ss* + *d*−1 *sn and δ* = 1 − (1 − *δs*)(1 − *δn*)*.*

## Simulation Setup

**Data preprocessing** We evaluate the performance on two

25: Compute *w*ˆ = arg min*w f*ˆ(*D , w*).

26: Compute *s* = 1 *ss* + *d*−1 *sn* and *δ* = 1−(1−*δs*)(1−*δn*).

datasets, *Adult* dataset and *US* dataset. The *Adult* dataset

from UCI Machine Learning Repository (Dheeru and

27: **return:** *w*ˆ, *s*

*d d*

and *δ*.

Karra Taniskidou 2017) contains information about 13 dif- ferent features (e.g., work-class, education, race, age, sex, and so on) of 48,842 individuals. The label is to predict

**Lemma 4.** *Let D and D* ∗ *be any two neighboring datasets differing in at most one tuple. Let f*¯(*D , w*) *and f*¯(*D* ∗*, w*) *be the approximate objective function on D and D* ∗*, then we have the following inequality,*

whether the annual income of those individuals is above 50K or not. The *US* dataset is from Integrated Public Use Micro- data Series (Center 2018) and consists of 370,000 records of census microdata, which includes features like age, sex, ed- ucation, family size, etc. The goal is to predict whether the

∗ ∗ ∗

. *d*2

income is over 25K a year. In both datasets, we consider sex

Δ2 = ǁ*A*1 − *A*2 ǁ2 ≤

*n*

+ 9*d.*

16

as a binary protected attribute.

**Baseline algorithms** In our experiments, we compare our

*where we denote A*1∗ = .Σ*i*=1 *λ*¯*φti* Σ*φ*∈∪2

*n*

*i*=1

*λ*¯*φt*∗

Φ *and A*2∗ =

approaches, PDFC, and ADFC against several baseline algo-

.Σ Σ

*i*

*φ*∈∪

2

*j*=1

Φ*j*

*as the set of polynomial coefﬁcients*

*j*=1 *j*

rithms, namely, LR and PFLR\*. LR is a logistic regression

*of f*¯(*D , w*) *and f*¯(*D* ∗*, w*)*. And we denote ti or t*∗*i as an ar-*

model. PFLR\* (Xu, Yuan, and Wu 2019) is a differentially

*bitrary tuple.*

Finally, by minimizing the differentially private objective

private and fair logistic regression model that injects Laplace noise with shifted mean to the objective function of logistic regression with fairness constraint. Moreover, we compare

**0.85**

**3)/5\***

**3')&**

**$')&**

**/5**

**0.8**

**0.75**

**$FFXUDF\**

**0.7**

**0.65**

**0.6**

**10-2 10-1 100 101**

**3ULYDF\ %XGJHW**

(a)

**0.76**

**0.74**

**3)/5\***

**3')&PDULWDO**

**3')&DJH**

**3')&UHODWLRQ**

**3')&UDFH**

**0.72**

**0.7**

**$FFXUDF\**

**0.68**

**0.66**

**0.64**

**0.62**

**10-2 10-1 100 101**

**3ULYDF\ %XGJHW**

(b)

**0.78**

**0.76**

**3)/5\***

**$')&PDULWDO**

**$')&DJH**

**$')&UHODWLRQ**

**$')&UDFH**

**0.74**

**0.72**

**$FFXUDF\**

**0.7**

**0.68**

**0.66**

**0.64**

**0.62**

**10-2 10-1 100 101**

**3ULYDF\ %XGJHW**

(c)

Figure 3: Compare accuracy under different privacy budgets on *Adult* (*δ* = 10−3).

our relaxed functional mechanism against the original func- tional mechanism proposed in (Zhang et al. 2012) and No- Privacy, which is the original functional mechanism without injecting any noise to the polynomial coefﬁcients.

**Evaluation** The utility of algorithms is measured by

*Accuracy*, deﬁned as follows,

*Number of correct predictions*

We can observe that ADFC continuously achieves better ac- curacy than PFLR\* in all privacy regime, and PDFC only outperforms PFLR\* when *s* is small. We also evaluate the effect of choosing different attributes as *xs* by performing experiments on *Adult* dataset. As shown in Figure 3b and Figure 3c, choosing different attributes, marital status, age, relation and race, has different effects on the accuracy of

*Accuracy* =

*,*

*Total number of predictions made*

PDFC and ADFC. However, PDFC and ADFC still outper- form PFLR\* under varying values of *s*. As expected, as the

which demonstrates the quality of a classiﬁer. The fairness of classiﬁcation models is qualiﬁed by *risk difference (RD)*

*RD* = | Pr(*y*ˆ = 1|*z* = 1) − Pr(*y*ˆ = 1|*z* = 0)|*,*

where *z* is the protected attribute. We consider a random 80- 20 training-testing split and conduct 10 independent runs of algorithms. We then record the mean values and stan- dard deviation values of *Accuracy* and *RD* on the testing dataset. For the parameters of differential privacy, we con- sider *s* = {10−2*,* 10−1*.*5*,* 10−1*,* 100*,* 100*.*5*,* 101}, and *δ* =

{10−3*,* 10−4*,* 10−5*,* 10−6*,* 10−7}.

## Results and Analysis

In Figure 1, we show the accuracy of each algorithm, func- tional mechanism, relaxed functional mechanism and No- Privacy, as a function of the privacy budget with ﬁxed *δ* = 10−3. We can see that the accuracy of No-Privacy remains unchanged for all values of *s*, as it does not provide any dif-

ferential privacy guarantee. Our relaxed functional mecha- nism exhibits quite higher accuracy than functional mech- anism in high privacy regime, and the accuracy of relaxed functional mechanism is the same as No-Privacy baseline when *s >* 10−1. Figure 2 studies the accuracy of each al-

gorithm under different values of *δ* with ﬁxed *s* = 10−2.

Relaxed functional mechanism incurs lower accuracy when

*δ* decreases, as a smaller *δ* requires a larger scale of noise to be injected in the objective function. But the accuracy of functional mechanism remains considerably lower than re- laxed functional mechanism in all cases.

Figure 3a studies the accuracy comparison among PFLR\*, LR, PDFC and ADFC on *Adult* dataset with the particular unprotected attribute *xs* denoted by marital status.

value of *s* increases, the accuracy of each algorithm becomes higher in above three ﬁgures.

Table 1 shows how different privacy budgets affect the risk difference of LR, PFLR\*, PDFC and ADFC on two datasets. Note that we consider the attribute *xs* as race on *Adult* dataset, and work on *US* dataset. It is clear that PDFC and ADFC produce less risk difference compared to PFLR\* in most cases of *s*. The key reason is that adding differ- ent amounts of noise regarding different attributes indeed reduces the correlation between unprotected attributes and protected attributes.

Table 1: Risk difference with different privacy budgets *s* on two datasets (*δ* = 10−3).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data | *s* | LR | PFLR\* | PDFC | ADFC |
| *Adult* | 0*.*01 | 0*.*187 ± 0*.*049 | 0*.*045 ± 0*.*095 | 0*.*048 ± 0*.*108 | 0*.*146 ± 0*.*131 |
| 0*.*1 | 0*.*187 ± 0*.*049 | 0*.*004 ± 0*.*009 | 0*.*005 ± 0*.*022 | 0*.*068 ± 0*.*028 |
| 1 | 0*.*187 ± 0*.*049 | 0*.*022 ± 0*.*088 | 0*.*002 ± 0*.*011 | 0*.*045 ± 0*.*027 |
| 10 | 0*.*187 ± 0*.*049 | 0*.*003 ± 0*.*001 | 0*.*035 ± 0*.*041 | 0*.*019 ± 0*.*003 |
| *US* | 0*.*01 | 0*.*191 ± 0*.*014 | 0*.*037 ± 0*.*038 | 0*.*003 ± 0*.*034 | 0*.*004 ± 0*.*007 |
| 0*.*1 | 0*.*191 ± 0*.*014 | 0*.*078 ± 0*.*021 | 0*.*001 ± 0*.*006 | 0*.*008 ± 0*.*003 |
| 1 | 0*.*191 ± 0*.*014 | 0*.*069 ± 0*.*007 | 0*.*022 ± 0*.*047 | 0*.*031 ± 0*.*004 |
| 10 | 0*.*191 ± 0*.*014 | 0*.*067 ± 0*.*003 | 0*.*022 ± 0*.*031 | 0*.*045 ± 0*.*002 |

# Conclusion

In this paper, we have introduced two approaches, PDFC and ADFC, to address the discrimination and privacy con- cerns in logistic regression classiﬁcation. Different from ex- isting techniques, in both approaches, we consider leverag- ing functional mechanism to the objective function with de- cision boundary fairness constraints, and adding noise with different magnitudes into the coefﬁcients of different at- tributes to further reduce the discrimination and improve the

privacy protection. Moreover, for ADFC, we utilize the pro- posed relaxed functional mechanism that is built upon Ex- tended Gaussian mechanism, to further improve the model accuracy. By performing extensive empirical comparisons with state-of-the-art methods for differentially private and fair classiﬁcation, we demonstrated the effectiveness of pro- posed approaches.

# Acknowledgments

The work of J. Ding, X. Zhang, and M. Pan was sup- ported in part by the U.S. National Science Foundation un- der grants US CNS-1350230 (CAREER), CNS-1646607,

CNS-1702850, and CNS-1801925. The work of X. Li was supported in part by the Programs of NSFC under Grant 61762030, in part by the Guangxi Natural Science Foun- dation under Grant 2018GXNSFDA281013, and in part by the Key Science and Technology Project of Guangxi under Grant AA18242021.

# References

Center, M. P. 2018. Integrated public use microdata series, international: Version 7.0.

Chaudhuri, K.; Monteleoni, C.; and Sarwate, A. D. 2011. Dif- ferentially private empirical risk minimization. *Journal of Ma- chine Learning Research* 12:1069–1109.

Datta, A.; Tschantz, M. C.; and Datta, A. 2015. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies* 2015(1):92–112.

Dheeru, D., and Karra Taniskidou, E. 2017. UCI machine learn- ing repository.

Ding, B.; Kulkarni, J.; and Yekhanin, S. 2017. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*.

Dwork, C., and Roth, A. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9(3–4):211–407.

Dwork, C.; Kenthapadi, K.; McSherry, F.; Mironov, I.; and Naor, M. 2006a. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006b. Calibrating noise to sensitivity in private data analysis. In *The- ory of cryptography conference*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*.

Ekstrand, M. D.; Joshaghani, R.; and Mehrpouyan, H. 2018. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*. Erlingsson, U´ .; Pihur, V.; and Korolova, A. 2014. Rappor:

Randomized aggregatable privacy-preserving ordinal response. In *21st ACM Conference on Computer and Communications Security*.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing dis- parate impact. In *Proceedings of the 21th ACM SIGKDD Inter- national Conference on Knowledge Discovery and Data Min- ing*.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model in- version attacks that exploit conﬁdence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*.

Giang, V. 2018. The potential hidden bias in automated hiring systems.

Hajian, S.; Domingo-Ferrer, J.; Monreale, A.; Pedreschi, D.; and Giannotti, F. 2015. Discrimination-and privacy-aware patterns. *Data Mining and Knowledge Discovery* 29(6):1733– 1782.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of oppor- tunity in supervised learning. In *Advances in neural informa- tion processing systems*.

Kamiran, F., and Calders, T. 2012. Data preprocessing tech- niques for classiﬁcation without discrimination. *Knowledge and Information Systems* 33(1):1–33.

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness- aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2012. A study of top- k measures for discrimination discovery. In *The 27th Annual ACM Symposium on Applied Computing*.

Pedreshi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination- aware data mining. In *14th ACM SIGKDD international con- ference on Knowledge discovery and data mining*.

Phan, N.; Vu, M. N.; Liu, Y.; Jin, R.; Dou, D.; Wu, X.; and Thai,

M. T. 2019. Heterogeneous gaussian mechanism: Preserving differential privacy in deep learning with provable robustness. *arXiv preprint arXiv:1906.01444*.

Rudin, W., et al. 1964. *Principles of mathematical analysis*. McGraw-hill New York.

Wachter-Boettcher, S. 2018. Ai recruiting tools do not elimi- nate bias.

Wang, Y.; Si, C.; and Wu, X. 2015. Regression model ﬁtting under differential privacy and model inversion attack. In *24th International Joint Conference on Artiﬁcial Intelligence*.

Xu, D.; Yuan, S.; and Wu, X. 2019. Achieving differential privacy and fairness in logistic regression. In *Companion Pro- ceedings of The 2019 World Wide Web Conference*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi,

K. P. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classiﬁcation without disparate mistreatment. In *The 26th International Conference on World Wide Web*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017b. Fairness constraints: Mechanisms for fair classiﬁcation. In *The 20th International Conference on Artiﬁcial Intelligence and Statistics*.

Zhang, J.; Zhang, Z.; Xiao, X.; Yang, Y.; and Winslett, M. 2012. Functional mechanism: regression analysis under differential privacy. *Proceedings of the VLDB Endowment* 5(11):1364– 1375.